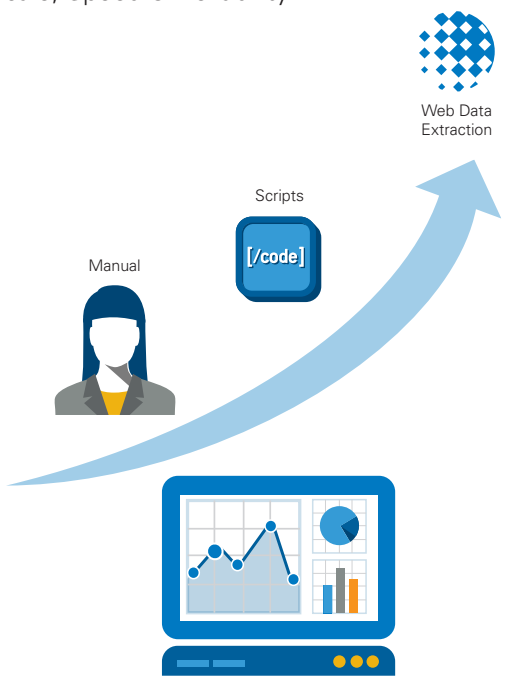
white paper

10 Must-Haves for Web Data Extraction and Transformation

A Guide to Harvesting, Extracting, and Integrating Web Data

KOFAX
from Lexmark

Power of Web Data Extraction:
Scale, Speed & Reliability



As data volume increases and the extracting process gets more complex, an enterprise-class web data extraction platform is needed.

Introduction:

To extract data from various internal and external sources is often a time-consuming process which involves either cutting and pasting data, using web scraping tools, developing custom homegrown scripts, or relying on applications that simply record a user's actions. None of these approaches can compete with the pace and ever-changing requirements of business. Over time, there will be an increased demand for quantity and quality of information.

Whether you are acquiring a million records a day from a few hundred sites or just a handful of records from thousands of sites daily, focusing on just one aspect of data handling will leave you short.

It's important to look at every aspect of how information is acquired, enhanced, and delivered into internal databases, spreadsheets, applications, and processes. There are significant differences among vendors and in-house custom solutions, and comparing them all can be a cumbersome chore.

Here's a start to defining the most important capabilities when it comes to extracting web data and transforming it be leveraged within any of your applications or processes.

Ten Must-Haves….

1. Coverage and integration with all web data sources

Data integration vendors use a variety of methods, technologies, and manual scripting practices to access the business data from disparate sources. The technology used often dictates your potential vendor's ability to harvest the quantity and quality of data you need to drive business decisions. Many vendors access only 60% (or less) of all web sites and cannot extract data from web sites that use AJAX, JavaScript, Flash, and PDFs. The impact is developers then have to write custom scripts to overcome the shortcomings of the product. Also note that not all web sites are created equally. While a vendor may be able to find a particular example of a Flash or AJAX site with which their tool works (possibly with quite a bit of customization or manual scripting prior to showing you the demo), they may not be able to access the Flash or AJAX site in which you're interested. The best approach to ensuring the product meets your needs is to provide the potential vendor with a list of the sites you want to access, not the ones they have prepared for you.

2. Automates web browsing navigation functions such as filling out forms and pagination (e.g., clicking to the "next" page).

As you discover the benefits of web data and become more dependent on real-time access, you'll need a product that extracts data in the same way a user will browse web sites. Specifically, this means filling in form fields, performing search queries, entering passwords, clicking through search results, choosing and comparing items, and moving from one page to another. Extracting the precise data you need from a web site, requires accuracy which can only be accomplished when a solution provides the ability to build robots that navigate sites much like a user would.

3. Surgically transforms unstructured web data to provide superior data quality without any errors or incomplete data.

While an automated data collection tool may appeal to you from an "ease of use" perspective, the real factor should be quality of transformed data. Without the ability to transform correctly, extracted data that is inaccurate or incomplete is useless. The downstream effect is significant in terms of cost and time spent manually cleaning up the data. The following capabilities should be considered important:

- Regular expressions: Search for text strings, including their variations (e.g., "grey" and "gray", "color" and "colour", or "car" and "cartoon"
- Encoding and decoding to deal with special characters
- Date formatting: handle international dates, convert time zones, deal with relative dates, combine dates, extract data "within the last 7 days" or "1 hour ago."
- String calculations
- Conditional expressions: if, then, else, and, or
- Numeric calculation: search a competitors price, compare to your own, find the difference, calculate the 10% difference, reduce your price 10%
- Multiple language support including multi-byte character sets

4. Provides automatic, customizable, and de-duplication of data.

The ability to automatically remove duplicate data will save you time, reduce confusion and eliminate manual preparation of the data. The more sites and data you extract, the more complicated and important this capability becomes.

5. Supports all data delivery output options.

Depending on where you are loading the data, you'll need to be able to output the extracted data into multiple destinations or formats, such as SQL database, MySQL, Java data structure, C# data structure, SOAP or REST web service, RSS, CSV or XML.

6. Automates all aspects of the data extraction and integration process.

Monitoring web sites, detecting changes, and extracting web data is a fairly complex task given web sites are built and configured differently, and >60% of dynamic sites are using JavaScript, AJAX, or HTML5. The most complete web data extraction and integration product will ensure 100% coverage, automation, and a more precise approach to the entire web. Furthermore, navigating multiple sites often requires using internal data to drive the interactions with a site, which requires bi-directional automation in order to extract the desired data. Do you have control over your data extraction integration flows, and can you quickly make changes, or does it require custom code changes and support from the vendor?

7. Supports different platforms and deployment models.

For Windows, Linux or UNIX support, be sure to verify your platform choice is supported. Furthermore, consider your preferred deployment mechanism and whether the vendor provides different options, including on premise or hosted cloud model, and the service level agreements and maintenance that go along with the various deployment options.

8. Scalability and performance.

The more sources and data you collect, the more important this question becomes. If you want to run multiple processes, does the product support load-balancing and multi-threaded execution to scale linearly with the number of CPUs? The use of headless browser technology will also be an important factor to reduce overhead and decrease strain on network bandwidth.

Being able to monitor and analyze real-time and historical process and system metric information is key to ensuring automated data extraction and integration processes are running smoothly and system performance is optimized. This is especially important for organizations collecting data from hundreds or even thousands of websites.

9. Debugging and error handling.

Can the product debug problems with automated extracting agents when they aren't working? Can you immediately see when a script or integration flow breaks, and do you have the ability to fix it on the fly? As web sites are dynamic and change constantly, how quickly you are be able to learn about the break and fix it is critically important. Look for a product that allows you to identify the breaks, apply a fix, and redeploy as quickly as possible to eliminate gaps in your data collection process.

10. Investment Cost and ROI.

The reality is low-cost web scraping and data extraction tools are not designed for large data extraction requirements that involve navigating complete websites and web portals, extracting precisely the data you want, and transforming into a usable format. At first glance, they look useful and easy to use, but upon closer inspection they involve scripting, are highly dependent upon the structure of a website, and simply cannot support large scale web data integration projects.

Now the Decision

As you move through an evaluation process, you'll want to dig deeply into the key topics and capabilities outlined above. This is especially important during the proof of concept phase. Make sure you're not opening up your wallet every time coding changes are required. You want to focus on specific capabilities noted above. In the end, you'll see a huge payoff in ROI if you zero in on the details and select the vendor that best addresses your needs.

And remember, web data extraction and integration is so much more than simply scraping data from a web site. Your business depends on the data. Whether it is being used to transform industries, grow market share, defend brands, or protect citizens, it takes a collective and precise approach to extract, transform and deliver the data you need, far beyond the reach of web scraping.

For more information visit kofax.com.

Use the checklist on the following page to evaluate web-scraping solutions

	KOFAX KAPOW™		VENDOR B		VENDOR C	
	YES	**NO**	YES	NO	YES	NO
Coverage and integration with of all web data sources	✔					
Automates web browsing navigation functions such as filling out forms, pagination	✔					
Ability to automatically transform unstructured web data	✔					
Provides automatic, customizable, and de-duplication of data	✔					
Supports all data delivery output options	✔					
Automates all aspects of the data extraction and integration process	✔					
Offers support for different platforms and deployment models	✔					
Scalability and performance	✔					
Debug and error handling	✔					
Resiliency and ability to adapt to changing web sources	✔					
Integration with business applications and processes	✔					
Ability to create and manage robotic integration flows	✔					
Ability to work with dynamic and unstructured web sources	✔					
Intelligent data extraction and transformation	✔					
Monitor and analyze web data integration process workflows and system performance	✔					

KOFAX from Lexmark

© 2015 Kofax. All rights reserved. Kofax, the Kofax logo, Kofax Kapow are trademarks or registered trademarks of Kofax in the U.S. and other countries. Other names may be trademarks or registered trademarks of their respective owners. (05.2015)